The role of duration as a cue for voicing of sibilants in Slovenian

Sašo Živanović[*], *Amanda Saksida*[**]

Abstract

We investigate the relative contributions of periodicity and duration as acoustic cues for voicing on sibilants in Slovenian. We find that (i) Slovenian exhibits the cross-linguistic tendency for voiced fricatives to be shorter than voiceless ones, and that (ii) periodicity and duration serve as complementary acoustic cues for identifying a sibilant as voiced or voiceless, with the relative contribution of the two cues varying across speakers.

Keywords: phonetics, fricatives, voicing, periodicity, duration

Vloga trajanja kot akustičnega namiga za zvenečnost sičnikov in šumnikov v slovenščini

V prispevku proučujeva relativni doprinos periodičnosti in trajanja kot akustičnih namigov za zvenečnost sičnikov in šumnikov v slovenščini. Ugotavljava, da (i) slovenščina sledi medjezikovni težnji, da so zveneči priporniki krajši od nezvenečih, ter da (ii) sta periodičnost in trajanje komplementarna akustična namiga za zvenečnost, katerih relativni doprinos razlikuje od govorca do govorca.

Ključne besede: fonetika, priporniki, zvenečnost, periodičnost, trajanje

* UL, Faculty of Arts, Ljubljana, Slovenia, saso.zivanovic@guest.arnes.si

** Educational research Institute, Ljubljana, Slovenia; Institute for Maternal and Child Health – IRCCS "Burlo Garofolo" – Trieste, Italy, amanda.saksida@gmail.com

1 Introduction

Voicing is a phonological feature which distinguishes classes of obstruents in many languages. On fricatives, it is acoustically realized primarily as a low frequency periodic component of the otherwise aperiodic high frequency signal. However, voiced and voiceless fricatives are also reported to differ in duration, with voiced segments tending to be shorter (see e.g. Cole and Cooper 1975; Baum and Blumstein 1987; Stevens et al. 1992; Nirgianaki et al. 2011; Jongman 2024).

In Slovenian, the durational difference between voiced and voiceless fricatives was only investigated by Jurgec (2019) and Kočevar (2024), who found that Slovenian follows the general tendency of voiced fricatives being shorter. However, investigating this difference was the primary goal of neither of these papers, leading to a less than optimal scope of the studies: Jurgec restricted his attention to stem-final segments, and furthermore investigated a specific, peripheral dialect of Šmartno, which exhibits several phonological properties absent in other dialects; Kočevar's subject were pre-school children, and she focused on the [s]/[z] pair and performed only a rudimentary statistical analysis. Furthermore, Jurgec's analysis did not include periodicity as an independent variable, while Kočevar's analysis calculated it using the arguably inferior centre of gravity method (see section 3.2).

The first aim of the study is to investigate whether the tendency of the voiced fricatives being shorter than voiceless ones is exhibited in Slovenian as well, using adult informants from non-peripheral dialects, inspecting the full set of phonological environments exhibiting the contrast between voiced and voiceless obstruents, and performing a thorough statistical analysis. We focus exclusively on sibilants as only these form lexical voiceless–voiced pairs ([s]–[z] and [ʃ]–[ʒ]) in Standard Slovenian (whereas [f] and [x] lack a lexical voiced counterpart), and also as they are easier to study due to a stronger overall amplitude.

While we expected that Slovenian would follow the general tendency described above, our preliminary observations also indicated that speakers differ with respect to which property of the signal, periodicity or duration, serves as the primary acoustic cue for voicing in their speech. The second, central goal of the study was therefore to see whether there is any inter-speaker variation with respect to the use of periodicity and duration as cues for phonological voicing, and whether this variation, if it exists, is categorical (in the sense that each speaker clearly favours either periodicity or duration) or gradual (in the sense that both cues contribute towards identifying a segment as voiced or voiceless, but that the relative contribution of the cues varies across speakers).

2 Methods

We are using the materials prepared, recorded and annotated by Teran (2020), who also performed a rudimentary statistical analysis of the data, under the mentorship of the first author. His materials are publicly available at osf.io/fz95h under the licence CC0 1.0 Universal. Our refined dataset is available (complete with processing scripts) at osf.io/s6zfd under the licence CC-By Attribution 4.0 International.

Participants: The study included eight adult participants (four male, four female) aged 20 to 65 from three dialect groups (one Lower Carniolan, two Upper Carniolan and five Styrian).[1] The participants were aware of the general scope of the study and gave their informed consent to participate. They were subsequently informed about the specific question of the study.

Materials: The participants read a short story specifically prepared to contain many occurrences of both voiced and voiceless sibilants in various phonological environments. In an attempt to make the informants' speech natural despite reading, they were encouraged to not use the standard language.

Recording equipment: The informants were recorded with a mobile phone, producing digital audio files in .wav format (mono, 44100 Hz). The low-end equipment is due to the fact that the stories were recorded for a BA thesis (Teran 2020) without a funding source. We are aware that it is customary in acoustic research to use professional-grade recording equipment in sound-proof booths. However, given the robustness of the speech signal (witness the human ability to understand speech under non-ideal conditions such as noisy environment, low bandwidth communication channel and overlapping sources), it is unlikely that this invalidates the analysis (cf. Benesty et al. 2008a; Droppo and Acero 2008).

Segmenting and annotation: The recordings were segmented and annotated using Praat (Boersma and Weenink 2025), see Figure 1 for an example. The beginning and end of each sibilant was marked on an interval tier. The sibilant boundaries were determined by the presence of the high-frequency noise characteristic for fricatives. Other portions of the recording were not further segmented, but they were fully transcribed on the same tier, marking segmental content, stress and morphosyntactic boundaries, thereby providing the information necessary to establish the phonological environment of each sibilant. We have also marked problematic spots such as

1 We include information on gender, age and the dialect for completeness. However, as expected, these variables turned out to not be statistically significant, so we will not discuss them in the analysis.

performance errors or intervals where external noise made the periodicity analysis impossible. Erroneous and error-adjacent sounds were ignored in the analysis.

The sibilant boundaries were determined by the presence of the high-frequency noise characteristic for fricatives. However, these boundaries are not exact. Segmentation depends on the gradual nature of the acoustic signal, phonological environment, speaker and random effects of the occurrence of the sound. Where clear delineation was impossible, the guideline was to segment counter to the expectation of voiceless sibilants being longer than voiced sibilants, i.e. to mark voiceless sibilants as shorter and voiced sibilants as longer, in order to prevent the influence of possible human bias in segmentation.

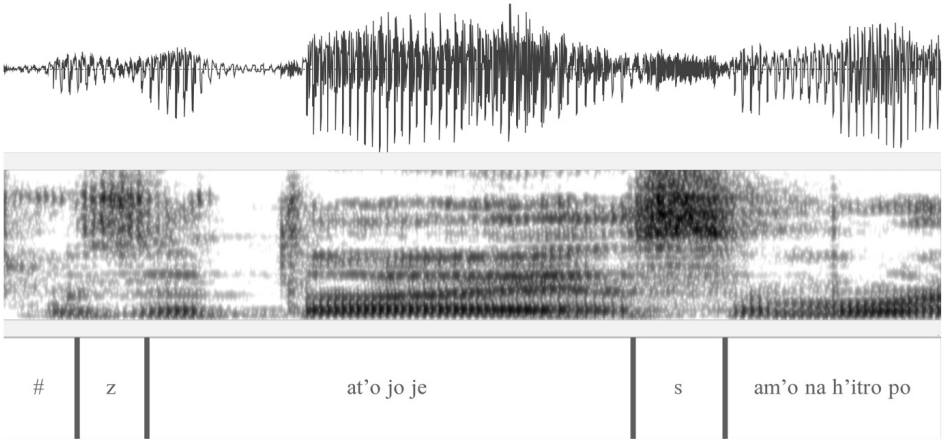


Figure 1: Annotation view showing (from top to bottom) the waveform, spectrogram and transcription. The spectrogram was drawn using Praat's default settings except that the frequency view range and dynamic range were set to 0–8000 Hz and 55 dB, respectively, for better visibility. The transcription shows that sibilants, but not other sounds, were segmented out. It includes a phrasal boundary (#) and stress marks (') immediately preceding vowels.

Preprocessing: We preprocessed the data using a Python script depending on packages Parselmouth (Jadoul 2024) and TextGridTools (Buschmeier and Włodarczak 2023), which offer a Python interface to Praat. The script converted the annotated recordings (i.e. the .wav and .TextGrid files) into a table (a .csv file) with columns informant *ID*, *segment* (s [s], š [ʃ], z [z], ž [ʒ]), phonological *environment*, segment *duration*, *periodicity*, *start time* and *end time*.

3 Measures

We deploy two gradual measures which we expect to correlate with the binary phonological category of voicing, one temporal and one spectral. The temporal measure is simply the duration of the segment. Voiceless sibilants (and fricatives in general) are expected to be generally longer than voiced sibilants. Periodicity is based on the spectral features of the acoustic signal. Voiced sibilants (and fricatives in general) are expected to have higher periodicity than voiceless sibilants.

3.1 Duration

The duration of each sibilant segment was computed as *end time* − *start time*, where *start time* and *end time* are the manually marked segment boundaries.

The computed durations are absolute, in (milli)seconds. We have tried several normalization methods, but as the results were comparable to those obtained with absolute duration, we ultimately decided against using normalization, both to keep the results easily interpretable and because the models using absolute values are better suited for a potential integration in sequential processing. The attempted normalization methods were the following:

- Speech-rate normalization, which is frequently deployed in analyses of connected speech (cf. e.g. Bjorndahl 2022; Jongman et al. 2000). The real duration of each recording was computed as the duration of the recording minus the manually marked pauses (an interval was interpreted as a pause if it contained only a boundary symbol and no real segment). Speech rate was then expressed as the mean syllable duration, computed as the real duration divided by the number of syllables (i.e. the number of vowels and syllabic consonants). Finally, the relative duration of a sibilant was computed as the ratio between the absolute duration and the mean syllable duration.
- Per-speaker normalization was supposed to account for individual differences in articulation of sibilants which go over and beyond differences in speech rate. The median duration of word-initial prevocalic occurrences of [s] in a stressed syllable was computed, per informant. The relative duration of a sibilant was then computed as the ratio between the absolute duration and this median.
- Normalization per phonological environment. For each environment (preceding a stressed vowel, unstressed vowel, sonorant or obstruent), we computed the median duration of all sibilants in the environment. The relative duration of a sibilant in an environment was then computed as the absolute duration divided by the median of

that environment. Note that using duration normalized in this way did not yield notably better models even though the duration variance between the environments was statistically significant.

3.2 Periodicity

There are several measures based on the spectral features of the acoustic signal which can be reasonably expected to correlate with the phonological category of voicing, both in general and specifically for fricatives. For example, the measure called harmonicity or Harmonics-to-Noise Ratio (HNR) directly exploits a basic articulatory fact that voiced fricatives have two sound sources: a turbulent noise source due to the rapid flow of air through a constriction of the airway, characteristic of fricatives, and a periodic glottal source due to vocal fold vibration, characteristic of voiced sounds (Stevens 2000). Acoustically, these two sound sources are reflected as the aperiodic high-frequency and the periodic low-frequency part of the spectrum, and HNR compares the amount of energy contained in these parts. Specifically, it is defined as the logarithm of the ratio between the periodic and the aperiodic part of the signal (for details, see e.g. Boersma 1993).

Other measures include centre of gravity (see e.g. Maniwa et al. 2009), a weighted average of all frequencies in the spectrum (voiced segments, containing the low-frequency energy produced by vocal fold vibration, are expected to have a lower centre of gravity than voiceless sounds, where this energy is absent), intensity-based measures (used by e.g. Chang 2008) such as the ratio of the intensities of the consonant and the adjacent vowels (again, this measure is expected to be higher for voiced consonants due to the energy produced by the vocal fold vibration, cf. e.g. Ladefoged 2003), and pitch-based measures.

The latter class of measures is widely used in phonetic research. In particular, this holds for Praat's Voice Report, and this was used to investigate voicing in fricatives by Smith (2013), Davidson (2016), and Bjorndahl (2022), among others. The wide-spread deployment of the Voice Report and other pitch-based methods is not accidental: Gradoville (2011) compares the validity of ten measures (including measure variants) for fricative voicing based on pitch/pulse, harmonicity, intensity and centre of gravity, concluding that "Praat's internal pulse-based voice report and the low-frequency-to-total intensities ratio provide the best match for what can be observed in the spectrogram and auditorily" (Gradoville 2011, p. 71).

Pitch-based measures are defined on the fundamental frequency (F0) of the signal at each frame. Computing F0 is not trivial, and several F0 determination algorithms

exist. Jesus and Jackson (2008, p. 15) compared eight of these implemented by open source software, and found that Praat's algorithms (Boersma 1993) provided the most accurate fundamental frequency, while being a close second best for voicing decisions.

Given the widespread usage, ready availability and positive reviews, we also decided to deploy Praat's pitch detection algorithm. Specifically, we used the autocorrelation variation, recommended for cases involving vocal fold vibration in the Praat manual. We have, however, decided against using the often deployed Voice Report. For one, Voice Report is based on the pitch contour found by Praat's path finder algorithm, which works non-locally, by attempting to find the best path through the pitch candidates. Two, Voice Report outputs the fraction of unvoiced frames, and is therefore binary at the level of an individual frame, interpreting it as either voiced or unvoiced. We believe this is why Voice Report exhibited both the floor and the ceiling effect with our data. The periodicity measure we developed uses the raw results provided by Praat's pitch detection algorithm (i.e. it does not rely on the pitch contour), and avoids across-the-board attenuation effects by being gradual at the level of each frame, computing the periodicity of a segment as a mean of the periodicity of the frames rather than the percentage of voiced frames.

In more detail, the periodicity of a segment was computed to be a unitless quantity between 0 and 1, where 0 indicates a fully voiceless segment, and 1 indicates a fully voiced segment. It was estimated using an algorithm based on Praat's pitch analysis. Each recording was first processed using Praat's raw autocorrelation method to produce a pitch object (.Pitch). We have used the default arguments for raw autocorrelation, except the following:

- The *pitch floor* was lowered to 50 Hz (default: 75 Hz) to make sure that the pitch could be computed with the same settings for all informants.
- The *voicing threshold* was lowered to 0.25 (default: 0.45). This setting was found to better suit the analysis of fricatives (the default is geared towards measuring intonation and vocal fold vibration in the production of vowels).

The periodicity of a segment was defined as the mean periodicity of all frames of the segment (the frame duration of the pitch object was the default 10ms). Each pitch frame consists of an ordered list of candidates, each determined by frequency and strength (a number between 0 and 1, indicating the ratio between the value of the autocorrelation at the time lag corresponding to the frequency, and the global maximum of autocorrelation at time lag 0). A pitch frame may also contain a candidate of frequency 0 Hz and strength 0, which is interpreted as a "voiceless candidate", i.e. it represents the possibility that the frame is voiceless.

The periodicity of a frame was defined as the strength of the first voiced candidate with a frequency below a certain threshold; if there was no such frame, the periodicity was set to 0. We have limited the candidate frequency because the periodic part of the spectrum corresponding to the vocal fold vibration is of low frequency, comparable to that of the pitch. The threshold was set to 1.5 times the median pitch of the entire recording. The median pitch was calculated using Praat's "Get quantile" function with the quantile parameter set to 0.5 to yield the median. (We computed the median rather than mean because Praat easily marks some stretches as having a pitch much higher than the actual fundamental frequency. Using the median is a quick method of disregarding these wrongly assigned pitches.)

Table 1: The Likert scale used to grade the perception of periodicity of segments in isolation (columns 1 and 2), cross-tabulated with their normal perception in the context of other segments (columns 3 and 4).

Perceived in isolation as		Perceived normally as		
value	interpretation	voiceless	voiced	total
1	clearly voiceless	194	1	195
2	most likely voiceless, but not a perfect example	411	20	431
3	unclear	169	87	256
4	most likely voiced, but not a perfect example	11	155	166
5	clearly voiced	1	188	189
total		786	451	1,237

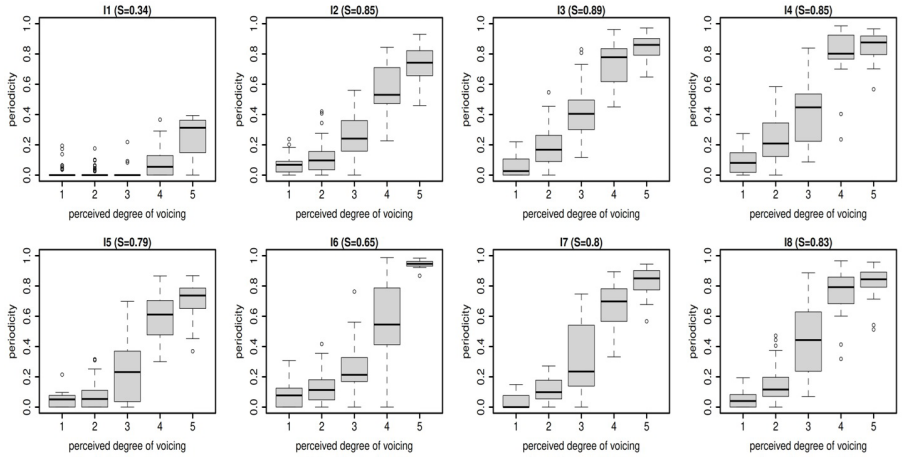


Figure 2: Correlation between the auditory perception of voicing in isolation (see Table 1) and the estimated periodicity, per informant. The plot titles include Spearman's rank correlation coefficient (*S*); the coefficient for the entire dataset is 0.75.

Several variations of the algorithm were tried out to find the parameter values, including the candidate frequency threshold factor of 1.5, and candidate selection method (only the first candidate counts) under which the estimates were found to largely agree with the authors' auditory perception of the amount of voicing of the segment played *in isolation* (to minimize the effect of phonological environment). To check the strength of this correlation, we annotated each target segment ([s], [ʃ], [z], [ʒ]) in the recordings with our auditory perception of the amount of voicing of the segment played in isolation.[2] The annotation results are shown in Table 1, and the resulting correlations with the computed periodicity are plotted in Figure 2 for each informant. The plot titles include Spearman's rank correlation coefficient of the correlation for each participant. We see that the correlation is (positive) strong for both for the entire dataset and for all informants, except informants I1 and I6, where it is weak and moderate, respectively.

4 Results

Preprocessing yielded 1,237 target segments (578 [s], 208 [ʃ], 301 [z], 150 [ʒ]), distributed across phonological environments as follows: preceding a stressed vowel 324, an unstressed vowel 429, a sonorant 157, an obstruent 327 (we disregarded phrase-final segments, as they are all voiceless due to final devoicing). The number of target segments produced per informant: 158, 150, 158, 156, 152, 155, 155, 153.

The box plots in Figure 3 show the values of segment duration and periodicity for the entire dataset. The voiced segments tend to be shorter ($t(1208.79) = -21.55$, $p < 0.001$) and more periodic ($t(617.12) = 29.76$, $p < 0.001$) than the voiceless segments, and the differences in duration and periodicity between places of articulation ([s]:[ʃ] and [z]:[ʒ]) are not statistically significant. Inspecting the scatter plot in Figure 4a (ignore the lines for now), one can see that duration and periodicity are moderately negatively correlated, the Pearson correlation coefficient is -0.45.

2 The annotations were performed individually by both authors. The inter-annotator agreement was almost perfect, with Cohen's kappa coefficient $\kappa = 0.86 \pm 0.01$, mean absolute score difference of 0.55 ($\sigma = 0.59$) and maximum absolute score difference of 2. Table 1 and Figure 2 show merged scores. Instead of averaging the scores, we took the score that was closer to the unclear score (3), e.g. we merged 1 and 2 into 2, and 3 and 4 into 3; in the rare cases where we scored on the opposite sides of unclear, we merged into the unclear score, i.e. we merged 2 and 4 into 3.

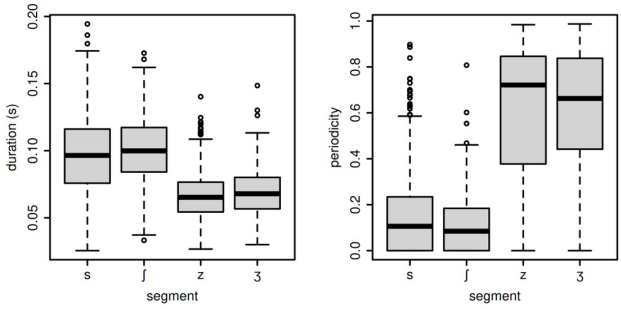


Figure 3: Box plots of duration (left) and periodicity (right) per segment.

The data was explored in detail by fitting Generalized Linear (Mixed Effects) Models (GLM(ER)),[3] which we evaluated using the following metrics: sensitivity (the rate of detecting voiced segments), specificity (the rate of detecting voiceless segments), F1 score, and area under curve (AUC). The latter metric is based on the receiver operating characteristic curve (ROC). The proportion of training data was 50%, and each test was iterated 1,000 times.

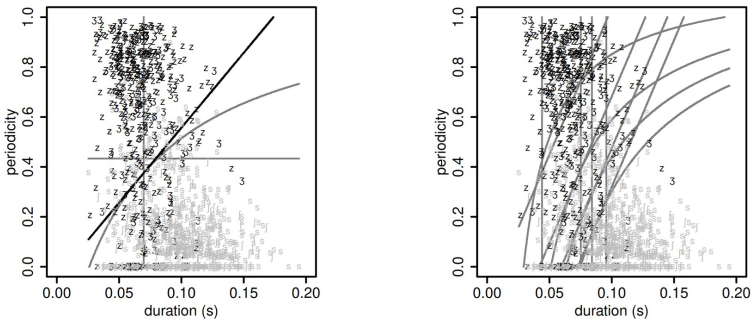
Table 2: Test scores for GLM models

formula	sensitivity	specificity	F1	AUC
$v \sim p$	0.74 ± 0.08	0.92 ± 0.05	0.79 ± 0.05	0.87 ± 0.04
$v \sim d$	0.59 ± 0.14	0.82 ± 0.07	0.62 ± 0.07	0.80 ± 0.04
$v \sim p + d$	0.75 ± 0.07	0.91 ± 0.04	0.79 ± 0.05	0.91 ± 0.03
$v \sim p * d$	0.76 ± 0.08	0.90 ± 0.06	0.79 ± 0.05	0.92 ± 0.03
$v \sim p + e$	0.74 ± 0.08	0.93 ± 0.04	0.80 ± 0.05	0.89 ± 0.04
$v \sim d + e$	0.74 ± 0.08	0.90 ± 0.05	0.77 ± 0.06	0.89 ± 0.03
$v \sim p + d + e$	0.80 ± 0.08	0.93 ± 0.04	0.83 ± 0.04	0.94 ± 0.02
$v \sim p * d + e$	0.82 ± 0.07	0.93 ± 0.04	0.84 ± 0.05	0.95 ± 0.02

Table 2 shows the model scores for models whose classifications are depicted in the scatter plots in Figure 4. Periodicity alone ($v \sim p$) is already quite predictive of voicing, while duration ($v \sim d$) is less so. Combining the measures, either with ($v \sim p$

3 We also performed the basic analysis using Support Vector Machines (SVM) and K-Nearest Neighbours (KNN) models. The differences were negligible. We only present the results obtained with the GML(ER) models, as they are the easiest to interpret and visualize.

$*$ *d*) or without ($v \sim p + d$) the interaction term, yields somewhat better performance, but as shown in the bottom part of the table, duration only truly develops its predictive power in tandem with phonological environment (*e*). (The environment term is also only effective in the presence of duration, cf. model $v \sim p + e$.)



(a) Horizontal line: v ~ p. Vertical line: v ~ d. Sloped line: v ~ p + d. Curve: v ~ p ∗ d.

(b) Vertical lines: v ~ d + e. Sloped lines: v ~ p + d + e. Curves: v ~ p ∗ d + e. Each line and curve of a parallel group belongs to the same model and corresponds to a phonological environment; from left to right: preceding an obstruent, an unstressed vowel, a sonorant, and a stressed vowel.

Figure 4: Scatter plot of duration vs. periodicity (the two plots are the same), with predictions of the GLM models. Voiced and voiceless segments are printed in black and grey, respectively. The dividing curves and lines show the classifications predicted by the GLM models at the cutoff probability of 50%. Voiced and voiceless segments are predicted to lie above and below the lines, respectively.

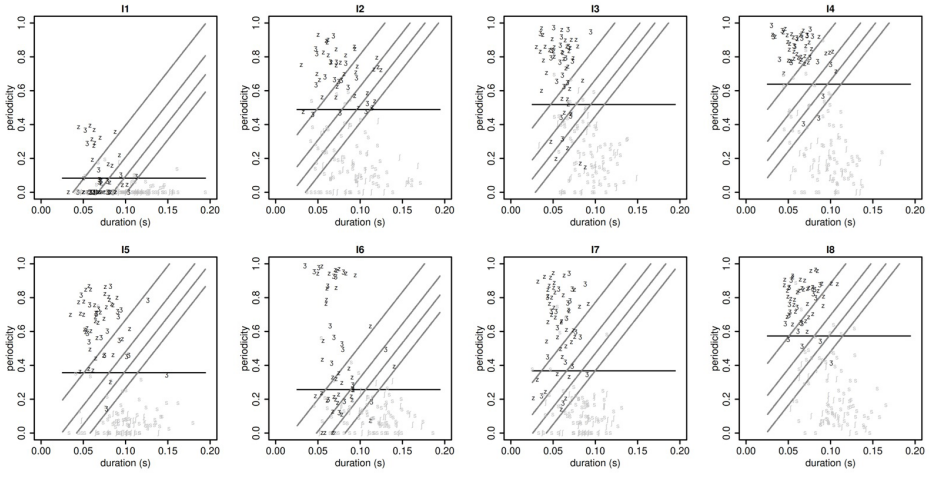


Figure 5: Scatter plot of duration and periodicity, with the predictions of the GLMER models $v \sim p + (1 \mid \text{ID})$ (black line) and $v \sim p + d + e + (1 \mid \text{ID})$ (grey lines).

Table 3: Test scores for GLMER models

formula	sensitivity	specificity	F1	AUC	success
$v \sim p + (1 \mid \text{ID})$	0.78 ± 0.09	0.94 ± 0.04	0.83 ± 0.05	0.95 ± 0.02	1000/1000
$v \sim p + d + e + (1 \mid \text{ID})$	0.88 ± 0.07	0.95 ± 0.04	0.89 ± 0.04	0.97 ± 0.02	1000/1000
$v \sim p + e + (d \mid \text{ID})$	0.88 ± 0.07	0.94 ± 0.04	0.89 ± 0.04	0.97 ± 0.01	775/1000

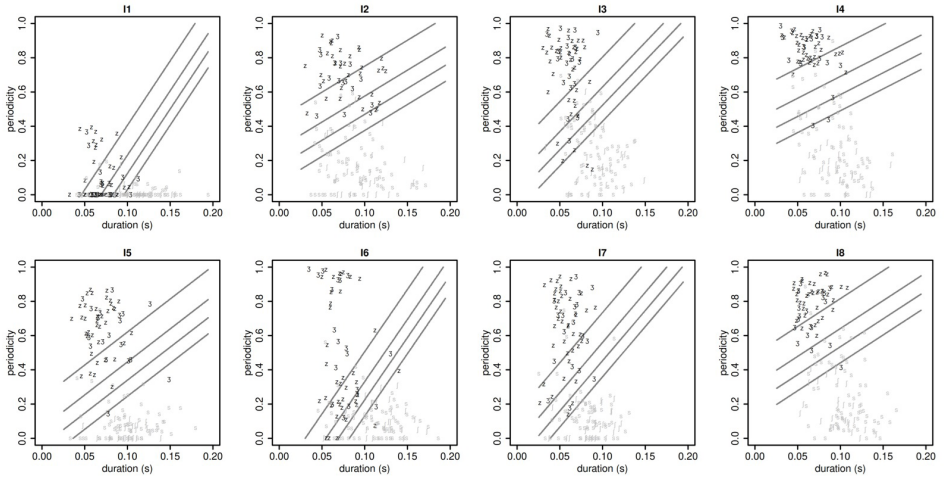


Figure 6: Scatter plot of duration and periodicity, with the predictions of the GLMER model $v \sim p + e + (d \mid \text{ID})$.

The above test results agree with those of the Type II ANOVA test performed on the full GLM model $v \sim p * d + e + g$: $\chi^2_p(1) = 306.66$ ($p < 0.001$), $\chi^2_d(1) = 206.11$ ($p < 0.001$), $\chi^2_e(3) = 195.18$ ($p < 0.001$), $\chi^2_g(1) = 3.68$ ($p = 0.06$), $\chi^2_{p:d}(1) = 17.38$ ($p < 0.001$). Note that although the interaction term $p : d$ is significant, its p-value (3.07×10^{-5}) is much larger than the p-values of other significant terms ($\leq 4.65 \times 10^{-42}$), which correlates with the fact that models with the interaction term ($p * d$) in Table 2 perform no better than the models without it ($p + d$). We therefore take the simpler and computationally less intensive $p + d$ variant as the basis for further models.

Our central research question is whether speakers exhibit individual differences in realization of voicing via duration, and whether such differences, if they exist, are categorical or gradual. Intra-speaker differences can be investigated using mixed-effect models with random effects for the informant.[4] Most often, we find researchers

4 In fact, not using mixed-effect models with random effects for the informant for our data would amount to violating the independence assumption.

use only random intercepts. In our case, this leads to model $v \sim p + d + e + (1 \mid \mathrm{ID})$ ($\chi^2_p(1) = 176.93$ ($p < 0.001$), $\chi^2_d(1) = 100.53$ ($p < 0.001$), $\chi^2_e(3) = 97.06$ ($p < 0.001$)) depicted in Figure 5 and evaluated in the upper half of Table 3. And while this is our best performing model – and incidentally, note that even the mixed-effects model $v \sim p + (1 \mid \mathrm{ID})$ featuring only the periodicity term performs excellently – it cannot answer our question, as it includes duration as a fixed term, which by definition cannot vary between individuals.

The model allowing the investigation of the intra-speaker differences in deploying duration must thus include duration as a random effect term. The random slope model $v \sim p + e + (d \mid \mathrm{ID})$ ($\chi^2_p(1) = 175.81$ ($p < 0.001$), $\chi^2_e(3) = 91.8$ ($p < 0.001$)) is shown in Figure 6 and evaluated in the lower part of Table 3.[5] The model arguably performs just as well as our best model $v \sim p + d + e + (1 \mid \mathrm{ID})$, even if the training does not always converge (see column "success"), presumably due to insufficient data.[6] The slopes of the classifier lines differ across informants, indicating intra-speaker differences in the use of duration for realization of voicing. If all classifier lines were (nearly) horizontal and vertical, corresponding to a small and large duration coefficient, respectively, the intra-speaker variation would be categorical. However, the observed minor variability of the slopes indicates gradual differences.

5 Discussion

The statistical analysis answered the questions posed in the introduction: Slovenian voiced sibilants are statistically significantly shorter than their voiceless counterparts, and speakers exhibit gradual differences in realization of voicing on sibilants. However, the most challenging part of the analysis was not the statistics, but implementing the measure of periodicity. The evaluation of the developed periodicity measure (see Figure 2) indicates that the measure is good but not perfect, with informant I1 standing out as the most critical. The question is whether one can define periodicity in a way which correlates even better with the perception data, and whether a better measure would improve the models even further (perhaps so that they could be used as a component of an automatic speech recognition system).

5 A lme4 GLMER model in R automatically includes the random intercept, i.e. d | ID is equivalent to 1 + d | ID.

6 Failure to converge is not necessarily indicative of a useless model, see https://www.rdocumentation.org/packages/lme4/versions/1.1-37/topics/convergence.

Figure 7 (ignore the model predictions for now) and Table 1 in section 2 offer a glimpse into how much a periodicity measure can possibly improve, and also hint at the general form the improvement should take. First, the cross-tabulation in the table makes it clear that there are limits to identifying segments in isolation: out of a 1,237 target segments, 1 + 20 + 11 + 1 = 33 segments (3%) were misheard and 169 + 87 = 256 (21%) segments remained unidentified. Second, mishearings were systematic: for a given speaker, the mishearings were either nearly all voiceless or nearly all voiced. Third, our periodicity measure is the least successful with the informants who were most often misheard (I1 and I6). The special status of these informants, and I1 in particular, is clearly visible when one compares the scatter plots in Figures 6 and 7. We interpret the situation as a signal that our periodicity measure works well for the majority of speakers, but that some speakers require a different approach – presumably because the spectral properties of voicing in their speech are different in some way. Summing up, we started the research with the hypothesis that there might be a categorical difference between speakers in terms of use of duration in the realization of voicing. The data showed that the intra-speaker difference is gradual rather than categorical. However, the above discussion makes one wonder whether the categorical difference exists after all, and if it is related to periodicity rather than duration.

To see whether an improved measure of periodicity would improve model predictions, we defined a fake measure of periodicity (p') partially based on the perception data (h). The idea is that this human-assisted measure should be near perfect in the sense of using all the available segment-internal spectral information, so that models fitted to the data using this measure could serve as the upper limit to what can be achieved by statistical analysis of the spectral and temporal information present in the acoustic signal of the segment. (Presumably, other linguistic modules and lexicon could fill in the remaining gaps in recognition.)

We did not use the perception data as the fake measure directly. This (categorical) data is valued on a Likert scale from 1 to 5, and while we could linearly translate the Likert values to the interval [0, 1] by $p' = (h - 1) / 4$, this would produce a discrete measure. We therefore decided to define a measure based on both the perception data h and our computed periodicity p. One option would be to simply multiply them ($p' = ph'$, where $h' = (h - 1) / 4$), but this would give too much weight to p; for example, a segment judged clearly voiced ($h' = 1$) but computed to be completely aperiodic ($p = 0$) would have $p' = 0$. We defined the measure in a step-wise fashion, as $p' = (h + 1 + p) / 5$: term $(h + 1) / 5$ zooms into one of the five equal-length intervals partitioning [0, 1], and term $p / 5$ adds some variability inside this smaller interval.

The fake periodicity measure was used to produce the models depicted in Figure 7 and evaluated in Table 4. Comparing Tables 3 and 4, we observe the slightly improved performance of all models (the simplest model $v \sim p + (1 \mid \mathrm{ID})$ improves the most, and sensitivity is the score with the highest increase). It is not surprising that the improvements are modest, as the models also fitted the real data extremely well, and the scores of model $v \sim p + d + e + (1 \mid \mathrm{ID})$ do indeed seem to be the reasonable upper limit we were trying to estimate.

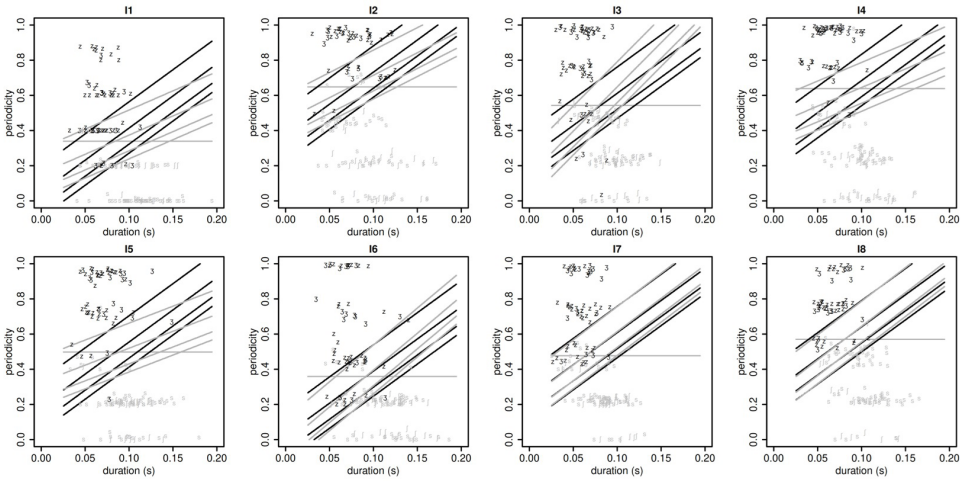


Figure 7: Scatter plot of duration and *fake* periodicity, with the predictions of the GLMER models $v \sim p + (1 \mid \mathrm{ID})$ (horizontal grey line), $v \sim p + d + e + (1 \mid \mathrm{ID})$ (black lines, parallel across informants), and $v \sim p + e + (d \mid \mathrm{ID})$ (diagonal grey lines).

Table 4: Test scores for GLMER models with the fake periodicity measure

formula	sensitivity	specificity	F1	AUC	success
$v \sim p + (1 \mid \mathrm{ID})$	0.88 ± 0.08	0.95 ± 0.04	0.90 ± 0.04	0.98 ± 0.01	995/1000
$v \sim p + d + e + (1 \mid \mathrm{ID})$	0.92 ± 0.06	0.96 ± 0.03	0.93 ± 0.03	0.99 ± 0.01	986/1000
$v \sim p + e + (d \mid \mathrm{ID})$	0.92 ± 0.06	0.97 ± 0.03	0.93 ± 0.03	0.98 ± 0.01	725/1000

6 Conclusion

Slovenian voiced sibilants are statistically significantly shorter than their voiceless counterparts. The duration of sibilants is moderately negatively correlated to their periodicity, a spectral measure of voicing. Periodicity is an excellent predictor of voicing, especially in mixed-effect models with a random speaker intercept. However, the periodicity–duration correlation is weak enough that duration can act as a redundancy measure. Adding a duration term improves model performance, especially when accompanied by the phonological environment term, which has no effect in the absence of duration. Inspecting and evaluating the models we found that periodicity and duration serve as complementary acoustic cues for identifying a sibilant as voiced or voiceless, with the relative contribution of the two cues varying across speakers.

References

Baum, Shari R. and Sheila E. Blumstein. 1987. Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. In: *The Journal of the Acoustical Society of America* 82, pp. 1073–1077.

Benesty, Jacob, M. Mohan Sondhi, and Yiteng Huang. 2008a. Introduction to Speech Processing. In: *Springer Handbook of Speech Processing*. Ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang.

Benesty, Jacob, M. Mohan Sondhi, and Yiteng Huang, eds. 2008b. *Springer Handbook of Speech Processing*.

Bjorndahl, Christina. 2022. Voicing and frication at the phonetics-phonology interface. An acoustic study of Greek, Serbian, Russian, and English. In: *Journal of Phonetics* 92, 101136.

Boersma, Paul. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: *Proceedings of the institute of phonetic sciences*. Vol. 17. 1193. Institute of Phonetic Sciences (IFA). Amsterdam, pp. 97–110.

Boersma, Paul and David Weenink. 2025. *Praat*. Version 6.4.26. URL: http://praat.org.

Buschmeier, Hendrik and Marcin Włodarczak. 2023. *TextGridTools (tgt)*. Version 1.5. URL: https://pypi.org/project/tgt.

Chang, Charles B. 2008. Variation in palatal production in Buenos Aires Spanish. In: *UC Berkeley PhonLab Annual Report* 4.4.

Cole, Ronald A. and William E. Cooper. 1975. Perception of voicing in English affricates and fricatives. In: *The journal of the acoustical society of America* 58.6, pp. 1280–1287.

Davidson, Lisa. 2016. Variability in the implementation of voicing in American English obstruents. In: *Journal of Phonetics* 54, pp. 35–50.

Droppo, Jasha and Alex Acero. 2008. Environmental Robustness. In: *Springer Handbook of Speech Processing*. Ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang.

Gradoville, Michael Stephen. 2011. Validity in measurements of fricative voicing. Evidence from Argentine Spanish. In: *Selected proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*. Ed. by Scott M. Alvord. Cascadilla Proceedings Project. Somerville, MA, pp. 59–74.

Jadoul, Yannick. 2024. *Parselmouth*. Version 0.4.5. URL: https://pypi.org/project/praat-parselmouth.

Jesus, Luis M. T. and Philip J. B. Jackson. 2008. Frication and Voicing Classification. In: *Computational Processing of the Portuguese Language. PROPOR 2008*. Ed. by António Teixeira et al., pp. 11–20.

Jongman, Allard. 2024. *Phonetics of Fricatives*. DOI: 10.1093/acrefore/9780199384655.013.1086.

Jongman, Allard, Ratree Wayland, and Serena Wong. 2000. Acoustic characteristics of English fricatives. In: *The Journal of the Acoustical Society of America* 108.3, pp. 1252–1263.

Jurgec, Peter. 2019. Opacity in Šmartno Slovenian. In: *Phonology* 36, pp. 265–301.

Kočevar, Eva. 2024. Analiza glasov /s/ in /z/ pri petletnikih. MA thesis. Univerza v Ljubljani, Pedagoška fakulteta.

Ladefoged, Peter. 2003. *Phonetic data analysis. An introduction to fieldwork and instrumental techniques*. Oxford: Blackwell.

Maniwa, Kazumi, Allard Jongman, and Travis Wade. 2009. Acoustic characteristics of clearly spoken English fricatives. In: *The Journal of the Acoustical Society of America* 125.6, pp. 3962–3973.

Nirgianaki, Elina, Anthi Chaida, and Marios Fourakis. 2011. Temporal characteristics of Greek fricatives. In: *Proceedings of the 9th International Conference on Greek Linguistics. 29–31 October 2009, University of Chicago, Chicago, Illinois*. Ed. by Katerina Chatzopoulou, Alexandra Ioannidou, and Suwon Yoon. Ohio State University, pp. 25–33.

Smith, Bridget. 2013. An acoustic analysis of voicing in American English dental fricatives. In: *Ohio State University Working Papers in Linguistics* 60, pp. 117–128.

Stevens, Kenneth N. 2000. *Acoustic phonetics*. MIT press.

Stevens, Kenneth N. et al. 1992. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. In: *The Journal of the Acoustical Society of America* 91.5, pp. 2979–3000.

Teran, Bine. 2020. Akustični izraz zvenečnosti na pripornikih v slovenščini. BA thesis. Filozofska fakulteta, Univerza v Ljubljani.